GTU - Eine Grammatik-Testumgebung

Jung, M ; Richarz, D ; Volk, Martin

Abstract: Die Grammatik-Testumgebung, GTU, bietet unter einer fensterorientierten Benutzeroberfläche ein integriertes Anwendungspaket, mit dessen Hilfe Grammatiken in drei Formalismen entwickelt und getestet werden können. Zum Testen steht eine umfangreiche Testsatzsammlung sowie ein uniformes und formalismusunabhängiges Lexikon zur Verfügung, das über ein integriertes Lexikoninterface an die Grammatiken angebunden werden kann. — In this paper we present a grammar development and testing tool called GTU (Grammatik-Testumgebung). GTU offers a window-oriented user interface that allows the development and testing of grammars under three formalisms. In particular it contains a collection of German test sentences and a uniform and independent lexicon, which can be adapted to a given grammar via an integrated lexicon interface.

GTU - Eine Grammatik Testumgebung

Michael Jung, Dirk Richarz und Martin Volk

Universität Koblenz-Landau, Institut für Computerlinguistik
Rheinau 1, D-56075 Koblenz
volk@informatik.uni-koblenz.de

Zusammenfassung. Die Grammatik-Testumgebung, GTU, bietet unter einer fensterorientierten Benutzeroberfläche ein integriertes Anwendungspaket, mit dessen Hilfe Grammatiken in drei Formalismen entwickelt und getestet werden können. Zum Testen steht eine umfangreiche Testsatzsammlung sowie ein uniformes und formalismusunabhängiges Lexikon zur Verfügung, das über ein integriertes Lexikoninterface an die Grammatiken angebunden werden kann.

Abstract. In this paper we present a grammar development and testing tool called GTU (Grammatik-Testumgebung). GTU offers a window-oriented user interface that allows the development and testing of grammars under three formalisms. In particular it contains a collection of German test sentences and a uniform and independent lexicon, which can be adapted to a given grammar via an integrated lexicon interface.

1 Einleitung

In den letzten Jahren hat sich der Computer zu einem wichtigen Werkzeug zur Syntaxanalyse natürlicher Sprachen entwickelt. Dies gilt für das Entwickeln, Testen und Vergleichen von Grammatiken, insbesondere im Hinblick auf das aufwendige Erstellen eines Lexikons, die Erzeugung von übersichtlichen Ausgabestrukturen sowie das Zusammenstellen von Testsätzen, die ein syntaktisches Phänomen in der jeweiligen Sprache ausreichend beschreiben.

Die bekannten Entwicklungswerkzeuge für Grammatiken unterstützen entweder einen bestimmten Grammatikformalismus (s. TAGDevEnv [6] für TAGs oder Alvey-GDE [3] für GPSG) oder sie lassen die Entwicklung beliebiger Grammatiken innerhalb des unifikationsbasierten Paradigmas zu (s. Pleuk [2] oder ELU [4]). Demgegenüber unterstützt GTU mehrere Grammatikformalismen (DCG, LFG und ID/LP) und ermöglicht dadurch ihren Vergleich. Das Lexikon kann über ein Lexikoninterface an die Bedürfnisse einer Grammatik angepaßt werden.

Soll eine Entwicklungsumgebung das Experimentieren mit Grammatiken fördern, so muß eine robuste und ergonomische Benutzeroberfläche gegeben sein. Ähnlich wie Pleuk bietet auch GTU eine graphische Benutzeroberfläche (unter X-Windows), über die alle Komponenten (Lexikon, Testsatzsammlung und Grammatikverarbeitung) ansprechbar sind. Die integrierte Testsatzsammlung ist das besondere Merkmal von GTU. Beim Testen einer Grammatik können Sätze daraus bequem ausgewählt und geparst werden.

2 Funktionalität von GTU

Die folgenden Punkte beschreiben die wichtigsten Aspekte von GTU. Das Lexikon und die Testsatzsammlung werden in eigenen Abschnitten vorgestellt.

Modularisierung der Grammatik: GTU unterstützt die Aufteilung einer Grammatik auf mehrere Module. Auf diese Weise kann der Benutzer die Regeln z.B. für die NP-Syntax von der VP-Syntax getrennt bearbeiten. Solche Teilgrammatiken lassen sich unabhängig von der Testsatzsammlung durch eine manuelle Eingabe von entsprechenden Satzteilen austesten. Die Grammatikmodule liegen in einzelnen Dateien und können getrennt übersetzt werden. Sollen Teile der Gesamtgrammatik ausgeblendet werden, so können einzelne Grammatikregeln und auch einzelne Module gezielt entfernt werden.

Übersetzung der Grammatik: Die Übersetzung der Grammatikdateien erfolgt durch ein eigenständiges GTU-Modul, der beim Ladevorgang die Grammatikregeln entsprechend des gewählten Formalismusses interpretiert und in PROLOG-Klauseln umsetzt. Bei DCG- und LFG-Grammatiken wird dabei aus der eingelesenen Grammatik ein äquivalenter Top-Down-PROLOG-Parser generiert, für eine ID/LP-Grammatik werden PROLOG-Klauseln erzeugt, die ein spezieller Bottom-Up-Chart-Parser verarbeiten kann (vgl. [8]).

Statisches Prüfen der Grammatik: Zur Überprüfung der Grammatik vor ihrer Anwendung durch den Parser werden verschiedene Optionen angeboten. So können die geladenen Grammatikdateien auf ihre Vollständigkeit bzgl. der verwendeten Kategoriebezeichner und ihre strukturelle Integrität getestet werden. Letzteres umfaßt zur Zeit den Test auf Linksrekursion bei Grammatiken für Top-Down-Parser (DCG, LFG) sowie den Test auf zirkuläre LP-Regeln.

Hilfe bei der Grammatikentwicklung: Für alle GTU-Funktionen stehen im Hauptmenü Hilfetexte zur Verfügung. Insbesondere ist die Syntax der Grammatikregeln und die Wortarteneinteilung des Lexikons jederzeit einsehbar.

Integrierte Editoren: GTU stellt Editoren zur Bearbeitung und Erstellung von Grammatik- und Testsatzdateien zur Verfügung. Diese Editoren sind in das Gesamtsystem integriert und erlauben das direkte Übersetzen und Laden der entwickelten Grammatiken oder Testsatzklassen.

Visualisierung der Parsingergebnisse: Die Ausgabe einer Satzstruktur nach dem Parsen eines Satzes erfolgt automatisch. Für alle Formalismen wird eine übersichtliche Baumstruktur angeboten. Die zugehörige Merkmalstruktur wird in Abhängigkeit von dem gewählten Formalismus präsentiert. Bei DCG- und ID/LP-Grammatiken erfolgt die Ausgabe der Satzstruktur in eingerückter Form, wobei die Konstituenten mit allen in der Grammatik spezifizierten Merkmalen versehen sind. Bei LFG-Grammatiken erfolgt anstatt dessen die Ausgabe einer F-Struktur, wie sie in der Theorie vorgesehen ist. Wird eine Eingabekette durch die Grammatik nicht vollständig beschrieben, so werden die größten erkannten Teilstrukturen ausgegeben. Die Ausgaben können wahlweise in ein oder mehrere Fenster sowie im ANSI- oder LaTeX-Format in eine Datei erfolgen. Damit lassen sich verschiedene Ergebnisse auch für den gleichen Satz nebeneinander darstellen oder für spätere Parsingvorgänge zu Vergleichszwecken aufbewahren.

3 Testsatzsammlung

Die Testsatzsammmlung ist mit rund 350 Testsätzen in zwei Organisationsformen in GTU integriert. Einmal sind die Sätze in 15 Klassen eingeteilt (in je einer eigenen Datei). Die Einteilung erfolgte aufgrund syntaktischer Merkmale, die den Sätzen als Annotationen mitgegeben werden. So gibt es z.B. Testsatzklassen für NP-Syntax und PP-Syntax, für Relativsätze und *daß*-Sätze, für Fragesätze und Aussagesätze. Jede Klasse enthält sowohl grammatische als auch ungrammatische Sätze, letztere zur Erkennung von Übergenerierungen einer Grammatik.

Die alternative Testsatzsammlung besteht aus den gleichen Testsätzen, die jetzt aber als Blätter an einem Merkmalbaum hängen, der die Beziehungen zwischen den einzelnen Syntaxphänomenen repräsentiert (z.B. *daß*-Sätze als Untergruppe der Nebensätze). Der Benutzer kann den Baum traversieren und erhält zu jedem Knoten die entsprechenden Sätze. Werden mehrere Knoten ausgewählt, so bildet das System die Schnittmenge der dadurch definierten Sätze. Dadurch können Sätze mit speziellen Merkmalkombinationen gefunden werden (vgl. [7]).

Die Standardisierung der Testsätze ermöglicht vergleichende Untersuchungen zwischen verschiedenen Grammatikformalismen bezüglich Effizienz und Kürze des Regelwerks. Die Sammlung ist modular aufgebaut und erlaubt die einfache Erweiterung um Testsätze und -klassen unter Verwendung spezieller Editoren, die in der GTU-Oberfläche eingebunden sind.

Der Benutzer kann mit mehreren Testsatzklassen gleichzeitig arbeiten. Zum Testen einer Grammatik wird eine ganze Klasse oder einzeln selektierte Sätze an die Analysekomponente übergeben. Die Sätze werden segmentiert, schrittweise morphologisch analysiert und unter Bezug auf die geladene Grammatik zum Parsen übergeben. Die berechneten Satzstrukturen werden je nach Benutzereinstellung im gewünschten Format angezeigt.

4 Lexikonkomponente

Die Lexikonkomponente (vgl. [5]) hat zur Aufgabe, dem Entwickler einer Grammatik die aufwendige Arbeit der Erstellung eines eigenen Lexikons zu ersparen. Sie besteht aus einem uniformen Lexikon für alle Grammatikformalismen sowie einem Lexikoninterface als Schnittstelle zwischen Lexikon und Grammatik.

Das Lexikon besteht überwiegend aus Stammformeinträgen, die die morphologischen und funktionalen Merkmale der Wörter beschreiben. Neben den Stammformen gibt es für einige Wörter Vollformeinträge, wenn das Wort unflektierbar oder die Flexionsklasse sehr klein ist. In GTU wurden die geschlossenen Wortklassen wie Determina, Pronomen, Konjunktionen, Adverbien, Hilfs- und Modalverben als Vollformeinträge und die offenen Wortklassen, wie Substantive, Adjektive und Vollverben als Stammformeinträge realisiert.

Der Lexikonaufbau bleibt für den Grammatikschreiber verdeckt. Er erhält nur die Information darüber, welche Wortarten im Lexikon vergeben wurden, welche Merkmale für jede Wortart verfügbar sind, welche Wörter zu welcher Wortart im Lexikon stehen bzw. welche morphologische Information für eine

gegebene Wortform zur Verfügung gestellt wird. Der Grammatikschreiber muß über Lexikoninterface-Regeln festlegen, welche Information aus dem Lexikon er für seine Grammatik benötigt und wie diese Information formatiert sein soll. Damit kann er z.B. definieren, daß bestimmte Formen der Possessivpronomen in der Grammatik wie Determina zu behandeln sind.

Zur Bearbeitung der Lexikoninterface-Regeln beinhaltet GTU ein eigenständiges Interfacemodul, das als Präprozessor vor der syntaktischen Analyse eines Satzes arbeitet. Die Analyse für einen Satz erfolgt wortweise, wobei die für jedes Wort erzeugten Merkmallisten anhand benutzerdefinierter Regeln in lexikalische Einträge der angegebenen syntaktischen Kategorie umgesetzt werden.

5 Ausblick

GTU wurde an der Universität Koblenz erfolgreich zur Unterstützung der Lehre in Computerlinguistik eingesetzt. Die einfache Anbindung einer Grammatik an das Lexikon sowie die automatische Visualisierung der Parsingergebnisse erwiesen sich als Motivationsschub für die Studierenden. Folgende Erweiterungen sind zur Zeit in Arbeit. Das CELEX- Lexikon (vgl. [1]) mit ca. 50.000 Stammformen des Deutschen wird derzeit angebunden und soll das bisherige Lexikon ersetzen. Selektionsrestriktionen werden integriert und für den LFG-Formalismus wird eine Komponente entwickelt, die aus der generierten F-Struktur eine logische Form erstellt. Schließlich ist ein Modul zur Protokollierung und Auswertung von Testergebnissen in Planung.

Literatur

1. Baayen, R.H.; Piepenbrock, R.; van Rijn, H.: *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania. 1993.
2. Calder, J.; Humphreys, K.: *Pleuk Overview.* University of Edinburgh, Centre for Cognitive Science. Technical Report, 1993.
3. Carroll, J.; Briscoe, T.; Grover, C.: *A Development Environment for Large Natural Language Grammars.* Computer Laboratory, University of Cambridge. Technical Report. 1991.
4. Estival, D.: *ELU User Manual.* Universität Genf: ISSCO, Technical Report, 1990.
5. Ridder, H.: *Eine adaptierbare Lexikonkomponente für unifikationsbasierte Grammatiken.* Studienarbeit. Koblenz: Universität Koblenz-Landau. 1991.
6. Schifferer, K.: *TAGDevEnv. Eine Werkbank für TAGs* In: Bátori, I. et al.: Computerlinguistik und ihre theoretischen Grundlagen, Berlin: Springer Verlag, 1988.
7. Volk, M.: *Einsatz einer Testsatzsammlung im Grammar Engineering.* Dissertation. Koblenz: Universität Koblenz. 1994.
8. Weisweber, W.: *Ein Dominanz-Chart-Parser für generalisierte Phrasenstrukturgrammatiken.* (KIT-Report 45) TU Berlin. 1987.